2012-10-24

Spectral Cross Correlation as a Supervised Approach for the Analysis of Complex Raman Datasets: The Case of Nanoparticles in Biological Cells

Mark Keating
Technological University Dublin, mark.keating@mytudublin.ie

Franck Bonnier
Technological University Dublin, Franck.Bonnier@tudublin.ie

Hugh Byrne
Technological University Dublin, hugh.byrne@tudublin.ie

Recommended Citation

Keating, M. E., Bonnier, F., & Byrne, H. J. (2012). Spectral cross-correlation as a supervised approach for the analysis of complex Raman datasets: the case of nanoparticles in biological cells. The Analyst. Royal Society of Chemistry (RSC). doi:10.1039/c2an36169h

Spectral Cross Correlation as a Supervised Approach for the Analysis of Complex Raman Datasets: The Case of Nanoparticles in Biological Cells

Mark E. Keating*, Franck Bonnier, Hugh J. Byrne,

Focas Research Institute, Dublin Institute of Technology, Kevin Street, Dublin 8, Ireland,

Abstract

Spectral Cross-correlation is introduced as a methodology to identify the presence and subcellular distribution of nanoparticles in cells. Raman microscopy is employed to spectroscopically image biological cells previously exposed to polystyrene nanoparticles, as a model for the study of nano-bio interactions. The limitations of previously deployed strategies of K-means clustering analysis and principal component analysis are discussed and a novel methodology of Spectral Cross Correlation Analysis is introduced and compared with the performance of Classical Least Squares Analysis, in both unsupervised and supervised modes. The previous study demonstrated the feasibility of using Raman spectroscopy to map cells and identify polystyrene nanoparticles in a lipid rich environment, which is suggestive of the membrane rich endoplasmic reticulum. However, short comings in identification of all nanoparticle signatures in the cell using K-means clustering are apparent, as highlighted by principal component analysis of the identified clusters which demonstrates that K-means clustering does not identify all regions where spectral signatures of the nanoparticles are evident. Thus, two more sophisticated analytical approaches to the extraction of the nanoparticle signatures from the Raman spectral data sets, namely classical least squares analysis and cross correlation analysis, were employed and are demonstrated to improve the identification of spectroscopic signatures characteristic of polystyrene nanoparticles in a cellular environment. Additionally, to investigate the local biochemical environment in which the nanoparticles are

trafficked, a pure spectrum of 3-sn-phosphatidyl ethanolamine was cross correlated against the Raman data set, further suggesting the particles are indeed localized in a lipid rich environment. Furthermore, to demonstrate the robustness and versatility of the analysis method, a spectrum of pure RNA was used to demonstrate that a differentiation could be made between DNA of the nucleus and RNA of the nucleolus using the supervised spectral cross-correlation technique.

Keywords: Raman Spectroscopy, NanoParticles, Intracellular localisation, Multivariate analysis, Classical Least Squares Analysis, Spectral Cross Correlation Analysis

***Corresponding Author:** Mark.Keating@mydit.ie

Introduction

Nanotechnology is set to become the first trillion dollar industry in history, with predicted benefits which span a wide range of fields, including applications in site specific delivery of drugs in humans, to antimicrobial paint coatings and textile finishing, to advances in the electronics industry [1–6]. However, there are caveats associated with deploying these nanotechnologies which must be addressed before true realistic applications can be widely accepted and adopted as the norm.

It is widely known that nanomaterials, more specifically nanoparticles, possess a range of unique characteristics which in some ways dictate their usefulness and applicability in fields such as medical science. Properties such as increased surface to mass ratio result in an increased reactivity and associated novel optical properties result in new possibilities in diagnostic and theranostic imaging and delivery[7,8], while novel semi-conductor properties are applicable to the electronics industry[9]. However, these properties also potentially have negative implications, most importantly in terms of the potential impact of nanoparticle exposure on human health and the environment. Nanoparticles have been demonstrated to be taken up by cells *in vitro* and to elicit a toxic response while many reports exist of adverse toxic effects *in vivo*[10–15].

One of the challenges facing the nanotoxicology community is the detection and monitoring of the interaction mechanisms of nanoparticles in cells[16,17]. Currently, fluorescent microscopy is the most widely used and accessible method to study nanoparticle uptake and trafficking[18–23]. Necessarily, however, it relies on the use of inherently fluorescent or labelled compounds for visualization and monitoring of nanoparticles inside cells. Most nanoparticles are not intrinsically fluorescent, however, and it has been recently demonstrated that fluorescent

labels can be labile, and that the observation and distribution of intracellular fluorescence following nanoparticle exposure is not necessarily representative of the presence or distribution of nanoparticles in the cell[24]. While it is also possible to study the dynamics of nanoparticle trafficking using label free optical microscopic techniques such as dark field and differential interference contrast (DIC) microscopy, the techniques are mainly applicable to metal particles such as gold and silver.[25] Transmission electron microscopy (TEM) provides an additional method by which nanoparticles can be visualised in a cellular environment[26–28]. The high lateral resolution obtainable with TEM renders it an ideal method for visualising sub cellular organelles and uptake and interaction of nanoparticles. However, significant sample processing (fixing and ultramicrotoming) is required and only particles with sufficient electronic contrast to the cellular environment can be visualised[28,29]

Thus, a label-free technique is required which can ideally unambiguously identify the presence of the nanoparticles in the cells, their sub-cellular location, and their overall effect on the cellular metabolism. Raman spectroscopy is one such method which may provide an alternate to traditional approaches for studying the nanoparticle-biological interface. The technique provides not only a label free method to visualize how the nanoparticle behaves in a biological environment, but offers the potential to identify the local environment and simultaneously analyse the associated metabolic changes. To do this, one must combine Raman spectroscopy with analytical data mining approaches to extract the signatures associated with the nanoparticles but also to probe the environment the particles are localized in, and to correlate the exposure and subcellular interaction mechanisms with the metabolic changes.

Previous studies have indicated the potential of Raman as a label free method for studying biological processes. Examples include novel approaches for cervical cancer

diagnostics[30], to investigating the effects following exposure to human papilloma virus (HPV) infection[31], the effects of chemotherapeutic anticancer agents in cells[32,33], live cell analysis[34,35]and the toxic responses to single walled carbon nano-tubes (SWCNT), to name but a few[36].

Surface enhanced Raman scattering (SERS) is also a potential method to study the intracellular dynamics of nanoparticle trafficking and compartmentalisation[37,38]. However, only certain types of nanoparticle, such as gold and silver particles and nanoaggregates have the potential to generate SERS spectra, thus limiting the technique to the study of only a certain type of nanoparticles. Additionally the surface enhancement process and molecular specificity of the technique are not fully understood, which may lead to ambiguity in the understanding of cellular trafficking.

A more recent study indicated the ability of Raman spectroscopy to detect the presence of intracellular polystyrene nanoparticles[39]. Polystyrene was chosen as a model nanoparticle for the study as it is commercially available and regularly employed as a standard in nanotoxicology (particularly as a positive control in its aminated form). Furthermore, the conjugated styrene ring makes it a relatively strong Raman scatterer. However, while the identification is somewhat straight forward, the presence of overlapping peaks in both the polystyrene and cellular spectra (e.g. both cellular and polystyrene spectra exhibit a strong symmetric ring breathing peak at ~1004cm^{-1}) presents a challenging system with which to validate the effectivity of the experimental and data analysis techniques. K-means clustering analysis (KMCA) analysis was used to differentiate regions of the cell as well as to identify and localise the nanoparticles. Analysis of the local cellular environment of the detected nanoparticles was performed via a comparison between loadings obtained from principal component analysis (PCA) and pure

spectra of lipids and polystyrene nanoparticles. However, when the data was analyzed using PCA, it was noted that the clusters detected using KMCA failed to identify all regions which contained the spectral fingerprint corresponding to polystyrene in a biological environment. Furthermore, the average spectra of the cluster identified by KMCA, while containing features clearly characteristic of polystyrene, also contained spectral features of the neighbouring cellular environment. Analysis of the loading of the principal components provided a clearer differentiation of the nanoparticle contributions from the local cellular environment, but neither unsupervised technique provided an unambiguous localisation of the target species [39].

Other multivariate analytical approaches have also been applied in the field of Raman microspectroscopy of cells. In addition to KMCA, other clustering methods such as Fuzzy C means clustering (FCM) and hierarchal cluster analysis (HCA) have been used to separate the cellular Raman data into clusters and subsequently reshape the data into images[40,41]. However, as highlighted by Headegaad *et al.*, these approaches have their own limitations. In particular boundaries between sub-cellular features can often result in the addition of extra clusters with mixed spectral signatures. This addition can be overcome by increasing the number of clusters; however, this in turn can result in added complexity to interpretation and inaccuracies in regional seperation. Additionally, the reproducibility of these methods can also be questioned as the starting point for the centroid based KMCA and FCM is subjective[40].

PCA and vertex component analysis (VCA) have also been used to separate out distinct regions of the cell. With regards to PCA, separation is based on the variances between the spectra in the data set, the majority of the variance being described by the first three principal components[40]. Thus, the score values can be used to construct a composite image of the cell in which the biochemical contributions of each component are described by the corresponding

loadings plot. Unlike KMCA and FCM, PCA identifies quite accurately the boundaries between each feature. However, the images generated suffer from inferior contrast and in some instances interpretation may be difficult as biochemical features may be spread across different loadings.

VCA is another method which has been used for similar analytical purposes. In brief, VCA computes a linear combination of supposed pure component spectra which are termed endmember spectra. As described in Miljkovic *et al.*, the endmember spectra are acquired under the assumption that the most extreme data points in the dataset are representative of pure component spectra[41]. However, it has been pointed out that the endmembers generated are not truly representative of the pure component they describe in the data set and can often contain a mixture of biochemical constituents i.e. DNA and proteins[42]. While this is representative of the true nature of nucleic acids *in-situ*, it could lead to inaccuracies in interpretation.

The work presented here demonstrates the potential of a Spectral Cross Correlation Analysis (SCCA) for the analysis of Raman spectral datasets. The method is applied to the dataset of Dorney *et al.* [39], of polystyrene nanoparticles in A549 lung adenocarcinoma cells, and is thus compared with previous analyses by KMCA and PCA. The performance of SCCA is also compared to that of classical least squares analysis (CLSA), performed both in a supervised and unsupervised manner, which allows for a direct comparison between both approaches. SCCA utilises the spectrum of the target chemical component and cross correlates the spectrum with that of the complete Raman spectral dataset. The quantitative performance is demonstrated using simulated datasets and the potential is demonstrated by mapping the spatial profile of the polystyrene nanoparticles in the cells as well as other biochemical components of the cell, (RNA and lipids).

Experimental

Sample Preparation for Raman Imaging

A549 Cells were seeded at a density of 4 x 10^4 cells onto calcium fluoride (CaF_2) windows (Crystran Ltd., UK) for confocal Raman imaging. The cells were incubated for 24 hrs in Dulbecco's Modified Eagle's Medium (DMEM F12), supplemented with 10% foetal calf serum (FCS) and 1% L-Glutamine at 37°C, 5% CO_2. Following cell adherence, 2 mLs of medium containing 1x 10^{12} nanoparticles per mL were added to the cells. The cells and nanoparticles were incubated for 24hrs at 37°C and 5% CO_2. Following nanoparticle exposure, the cells were washed in warm PBS three times and fixed for 10mins in 10% buffered formalin. After fixation, the cells were washed to remove any trace of fixative and kept in NaCl solution prior to imaging.

Component spectra used in SCCA were generated as described in Bonnier and Byrne 2012[43]. For polystyrene nanoparticle spectra, nanoparticle suspension was added drop-wise to a CaF_2 window and allowed to air dry prior to Raman acquisition. RNA from baker's yeast (*saccharomyces cerevisiae*) was added to water and subsequently deposited on a CaF_2 window and allowed to air dry. 3-sn-phosphatidyl ethanolamine was dispersed in chloroform and deposited on CaF_2 windows.

Confocal Raman Spectroscopic Imaging

Confocal Raman Spectroscopic Imaging was performed using a Horiba Yobin-Yvon LabRAM HR800 spectrometer with a 785nm, 300mmW diode laser as source and a Peltier cooled 16-bit

CCD. A 100X, N.A. 1.2, (LUMplanF1, Olympus) water immersion objective was used for all cellular measurements. The confocal pin hole of the system was set to 100µm, the recommended setting for confocal operation, to allow optical sectioning of the sample. A 300 lines per mm spectroscopic grating, providing a dispersion of ~1.5cm^{-1} per pixel, was used and the system was pre-calibrated to the spectral line at 520.7cm^{-1} of silicon. Using an automated programmable stage, Raman spectra of the cell were acquired with a 0.75µm step size over a 29*39 pixel area which encompassed the nuclear, perinuclear and cytoplasmic regions of the cell.

Data Pre-Processing and Preparation

In order to prepare the data for analysis, a number of steps were taken to ensure the spectra in the map were of a high enough quality to give accurate results. For CLSA, all data pre-processing was carried out using Labspec 5 software which comes as standard on the Raman instrument. Firstly, a background spectrum which constituted the contribution of the CaF_2 substrate and water in the imaging medium was subtracted from each spectrum in the mapped data set. Following subtraction of the background spectrum, a Savitsky-Golay smoothing filter (5^{th} order, 7 points), available on the software, was used to lightly smooth the data. The data was then baseline corrected using a nodal point baseline correction using the minimum amount of points possible to ensure minimal alteration of the acquired data. Normalization was carried out automatically by the software during CLSA.

Data was prepared in a similar fashion for SCCA. However, the pre-processing was carried out in Matlab (Mathworks,USA) using previously published protocols for data processing[39]. As outlined above, a background spectrum was subtracted from the Raman data set to remove the substrate and immersion medium contributions. A Savitsky-Golay smoothing filter

(5^{th} order, 7 points) was applied to the data and a nodal point baseline correction was used to baseline the data using a minimum amount of reference points to do so. Preparation of component spectra for SCCA was done in the same manner for polystyrene, RNA and lipids.

Classical Least Squares Analysis

CLSA was carried out using Labspec 5 software which comes as standard on the Raman spectrometer software. The analysis method is based on a fit of a linear combination of reference component spectra to the spectra contained in the raw spectral map. This is described by Equation 1, for the case where three reference component spectra are used. S is the sum of the linear contribution of the reference components (A, B, C), and x, y, z are the respective weightings or scores necessary for the weighted sum of the reference component spectra to match the raw data.

$$S = [x*A] + [y*B] + [z*C] \qquad \text{Equation 1}$$

Using the software, there are two different ways to obtain the reference component spectra. The first way is to obtain a pure spectral reference from a compound or compounds which can then be fitted according to Equation 1. The second method uses a factor analysis algorithm to generate the component spectra, the weighted sum of which is compared to the Raman spectral data set. Using the latter of the two methods, Zavaleta et al demonstrated the power of the technique to quantify quantum dot accumulation in an *in-vivo* mouse model and to separate out the different spectral contributions from complex SERS signals in the same data set[44]. In a similar and different way, both approaches to CLSA are explored to extract spectra which contain polystyrene nanoparticles and define other biochemical regions such as the RNA and lipid rich

environments. The relative contributions of the different components are defined by the weighting factors (x, y, z….).

Spectral Cross Correlation Analysis

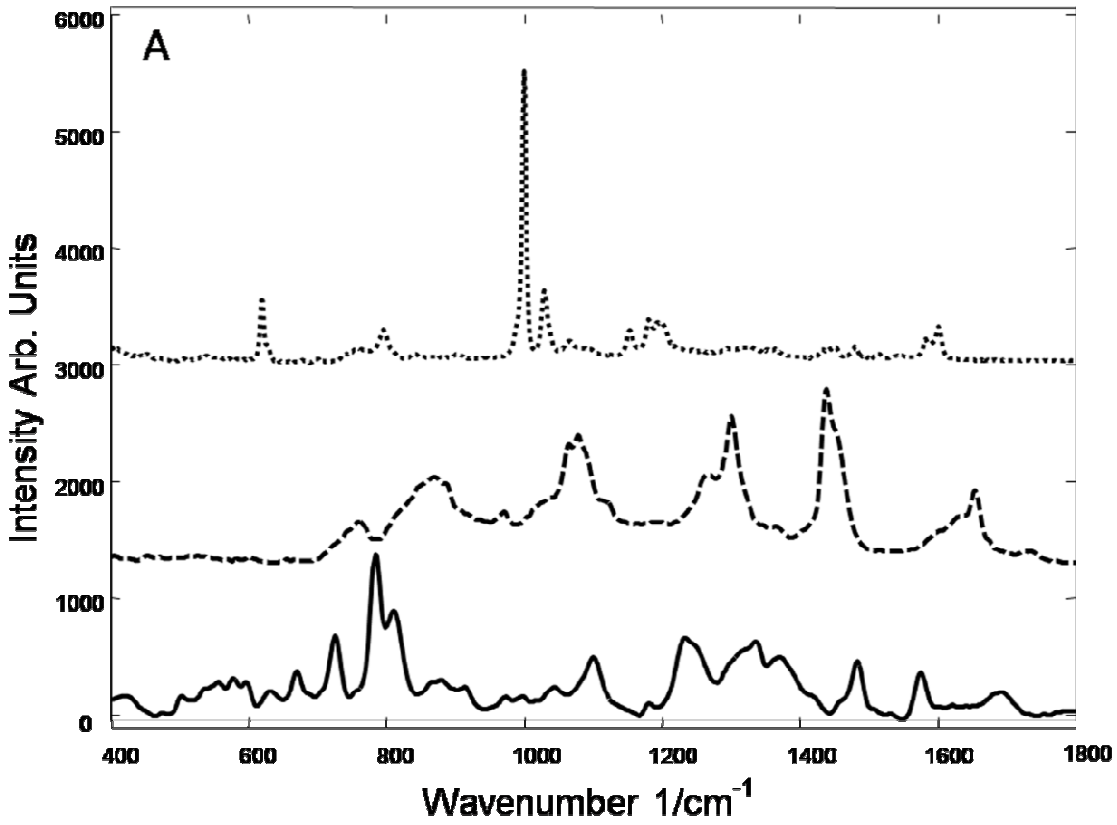
For SCCA, reference spectra from polystyrene, phosphatidyl-ethanolamine and RNA (Figure 1A) were used to screen the Raman spectral data set. All SCCA was carried out using Matlab (Mathworks, USA) using the "crosscorr" function available in the signal processing toolbox. Equation 2 describes the cross correlation between two data series, where C(x) is the correlation function, S(τ) is the Raman spectrum in the data set to be tested and A(x+τ) is the reference spectrum i.e. polystyrene, lipid or RNA. The function integrates the product of the two data series (spectra) at each point as they are shifted relative to each other along the x axis (wave number). The magnitude of the correlation quantifies the relative contribution of the component spectrum at that point in the cell, and an exact correlation occurs when the spectra are exactly matched (auto-correlation). In this way, it is possible to screen the map or spectra in the map and, based on the cross correlation function, cluster different biochemical regions of the cell based on the relative contributions of the reference spectrum used.

$$C(X) = \sum_{m=-\infty}^{\infty} S(\tau).A(X+\tau) \qquad \text{Equation 2}$$

Simulated Data

Simulated data sets were used to test the robustness and sensitivity of both CLSA and SCCA in their ability to detect spectral contributions due to polystyrene, RNA and lipid in a biological

environment. To generate the simulated data sets, a cellular spectrum was used as a template to which varied amounts of component spectrum were added. Keeping the cellular spectrum constant, a series of 38 simulated spectra of ratios 1:1 to $1:10^{-4}$, cellular: component Raman spectra for polystyrene, RNA and lipid were generated (Figure 1A). An example of the simulated data set for polystyrene is shown in Figure 1B, which shows the addition of the first 8 spectral dilutions to the constant cellular spectrum. Using these simulated datasets, it was possible to explore how each data mining approach performs when testing experimental data and thus facilitate accurate interpretation of the data sets.


A
Intensity Arb. Units
Wavenumber 1/cm-1
6000
5000
4000
3000
2000
1000
0
400
600
800
1000
1200
1400
1600
1800

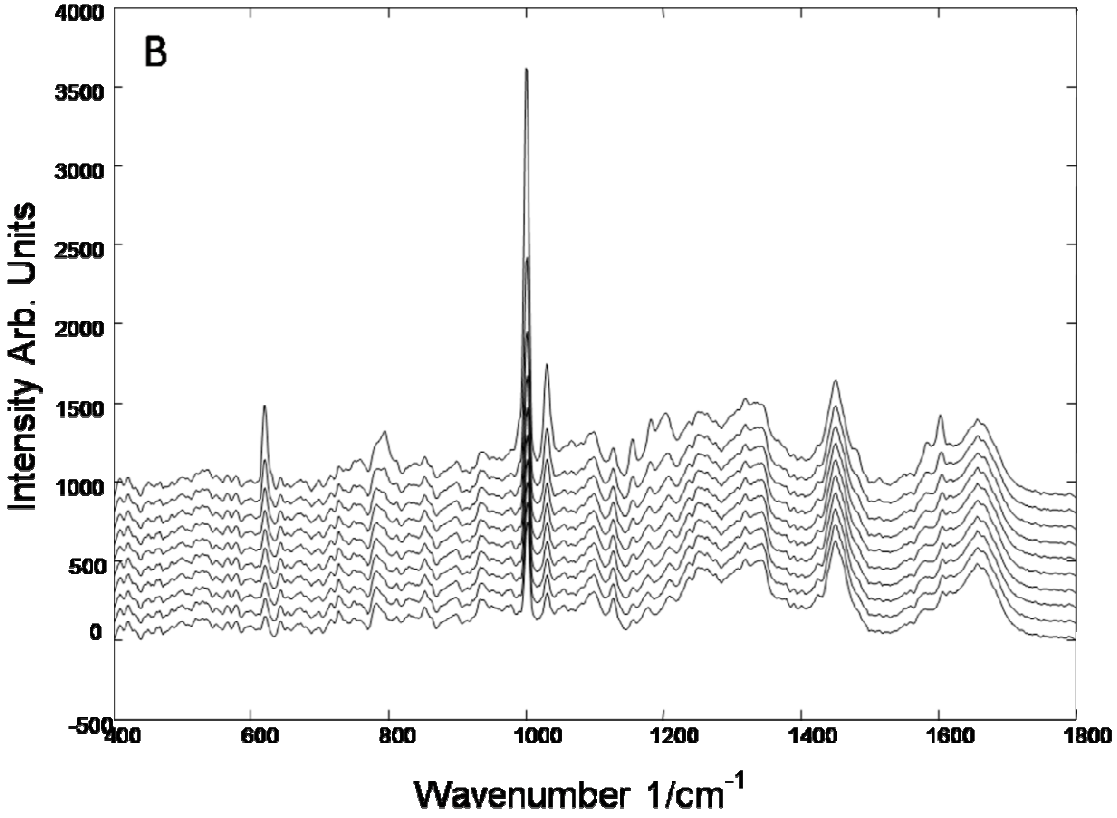

B
Intensity Arb. Units
Wavenumber 1/cm-1
4000
3500
3000
2500
2000
1500
1000
500
0
-500
400
600
800
1000
1200
1400
1600
1800

Figure 1. (A) Component spectra of nano-polystyrene (dotted line), 3-sn-phosphatidyl ethanolamine (dashed line) and isolated RNA (solid line), offset for clarity. (B) Shows an example of the first eight simulated spectra for polystyrene in cells, offset for clarity. Each spectrum consists of a constant cellular spectrum with a varied concentration of polystyrene added to it, with decreasing polystyrene concentration from top to bottom. Simulated data sets generated in this way were then analysed by CLSA and SCCA.

Results

Simulated Data – Unsupervised CLSA

CLSA can be carried out in two different ways, either by generating spectral models using a factor analysis algorithm (unsupervised), or by manually inputting the component spectra (supervised). The data in Figure 2 shows the results using the factor analysis generated models for simulated data sets generated based on cellular/polystyrene, RNA and lipid spectra (Figure 1 B). In each instance, the score recorded form CLSA for each spectrum is plotted against the component concentration added to the data set. In all cases, the extracted CLSA scores accurately represent the true component ratios over the concentration range, represented by the solid line. The results depart from nonlinearity a cellular:component ratio of ~1:0.1, after which the CLSA weightings no longer accurately reflect the correct component weighting, although the presence of the component can still be identified in ratios as low as 1:0.03.

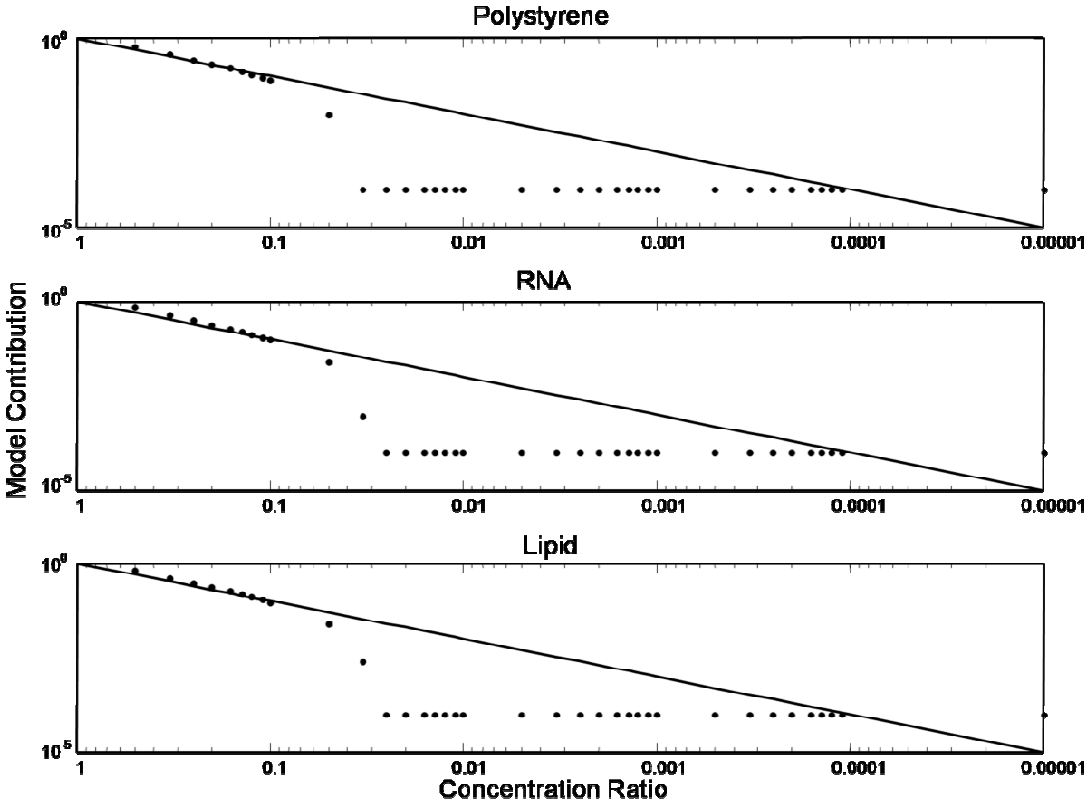


Figure 2. CLSA of simulated spectral data sets of nano-polystyrene, RNA and lipid. In each graph, the score from the CLSA is plotted against the concentration of component spectrum added to a constant cellular spectrum (points on each graph). The solid black line represents the ideal response which gives an indication of the quantitative nature of the technique.

Single Cell Data – Unsupervised CLSA

In order to further test the ability of CLSA to identify intracellular polystyrene nanoparticles located inside a single cell Raman map, an initial factor analysis algorithm was applied to the data set to generate 7 model spectra to be used in the CLSA. These model spectra were then used to compute the scores from the Raman data set (Figure 3A). It is then possible to segment the

cell into different distributions based on specific spectral differences as shown in Figure 3B. The spectral profile of each model contribution can be visualized individually showing the percentage contribution at each pixel (Figure 3C-F). A more detailed look at the model spectra generated and corresponding cellular distribution can be seen in Figure 4A-G.

The CLSA map shows a different spatial distribution of each model in the Raman spectral data set. Although in all cases, the model spectra show strong contributions of the cellular environment, they are differentiated by contributions from distinct components. Model 1 (Figure 4A) shows characteristic peaks corresponding to those seen in pure polystyrene spectra (see Figure 1A). Therefore, the pixel distribution of model 1 is deemed to show the localisation of the polystyrene nanoparticles, indicating a perinuclear distribution in the cell, consistent with the K-means cluster analysis of Dorney et al[39]. Other models show a different distribution in the cell. Model 6 shows a distribution which visually corresponds to the nucleolus of the cell (Figure 4B), whereas model 3 surrounds the nucleoli and is identified as the nucleus of the cell (Figure 4E). This shows the ability of CLSA to differentiate the biochemical regions of the cell containing RNA and DNA. Other models such as model 4 (Figure 4C) and model7 (Figure 4F) show a distinct distribution surrounding the nucleus, which may correspond to perinuclear organelles such as the endoplasmic reticulum or the Golgi apparatus which are lipid rich regions of the cell.

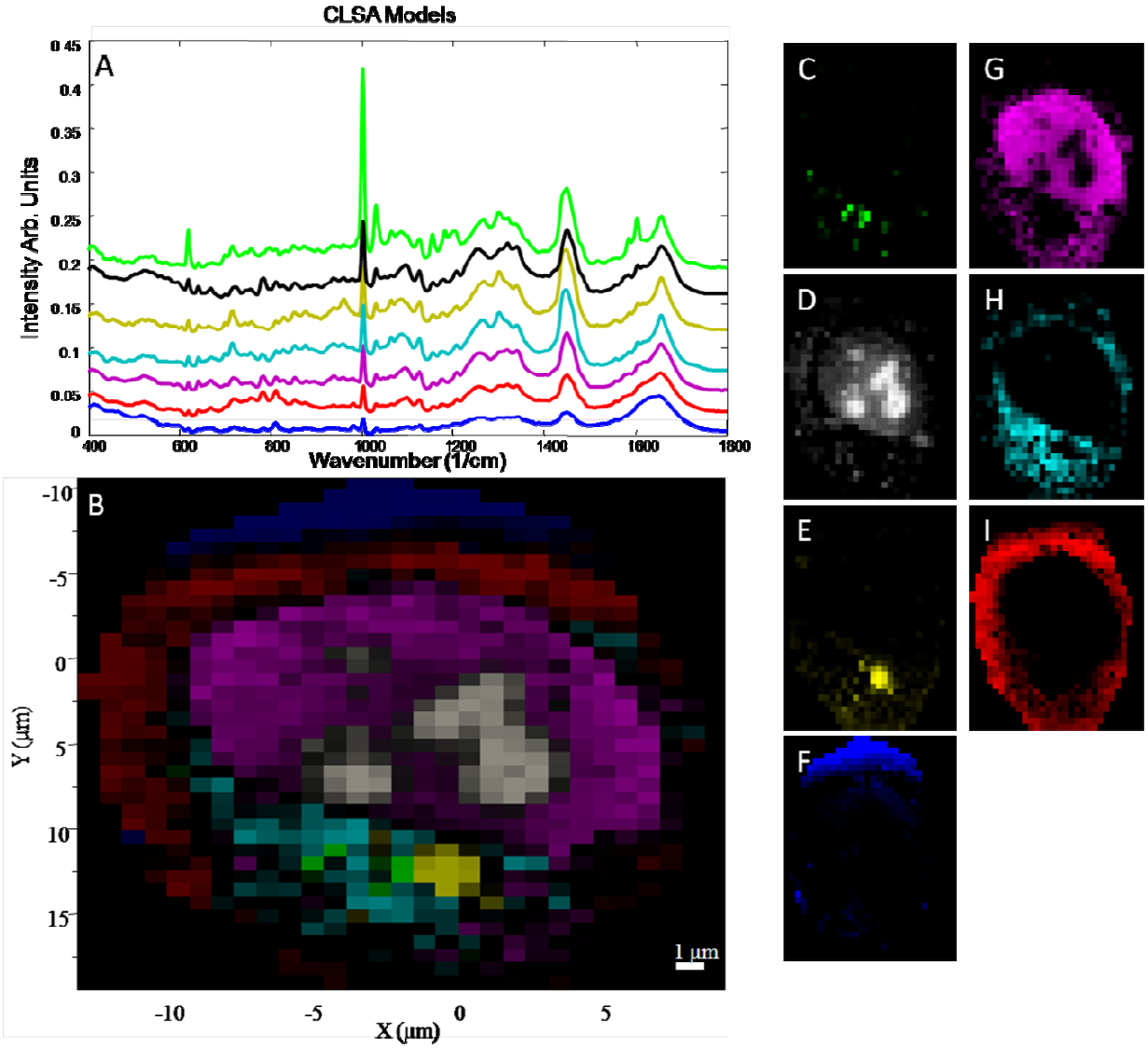


Figure 3: Clustering of spectra identified by unsupervised CLSA. (A) Spectral models generated from the analysis protocol and used to generate the clustered map shown in (B). The right panel (C-I) shows the distribution of each model created in the map. Of particular note, model 1(C), model 6(D) and model 7(H) have strong contributions of the spectra of polystyrene, RNA and lipid respectively. The spectra in (A) are colour coded and correspond to images (B – F), with the exception of Model 6 which corresponds to the white image in (D).

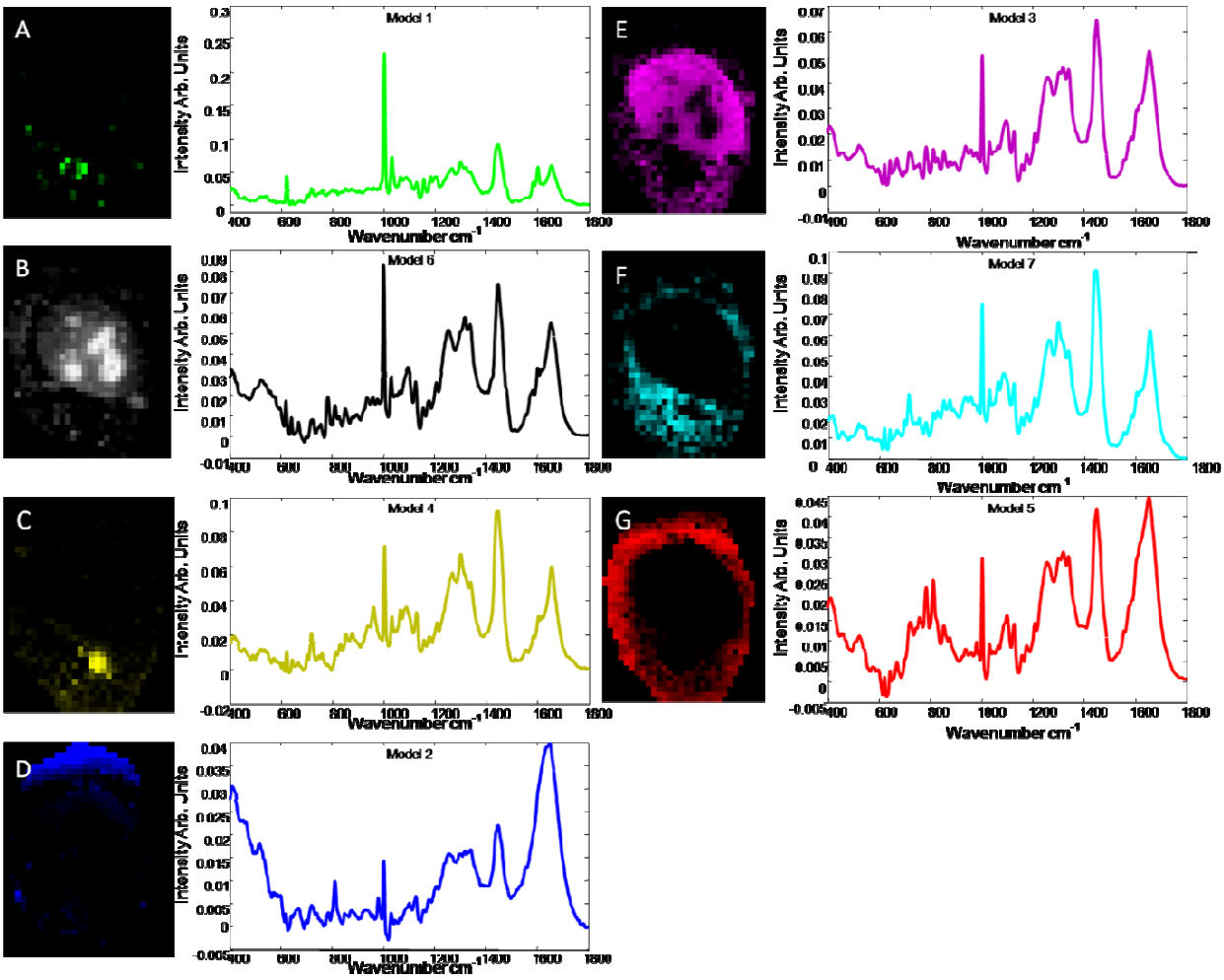


Figure 4: A closer look at the generated model spectra created by CLSA (A-G). The overlap between pixels corresponds to a percentage contribution from each particular model. In some instances a pixel may contain 50% of one model and 50% of another, which is highlighted somewhat by the intensity of the pixel, although this is visually subjective.

Simulated Data - Supervised CLSA

Unsupervised CLSA is clearly a powerful technique to analyse the subcellular structure and to identify the presence and distribution of nanoparticles. However, it should be noted that the technique does not yield pure spectra of the components (compare for example Figure 4A with

the pure spectrum of polystyrene in Figure 1A), and the respective models are mixtures of spectral signatures of the components and the background cellular spectrum. A secondary approach to CLSA which provides a more supervised approach was therefore also tested. In a similar way, the simulated datasets were used to assess the technique prior to testing the real Raman cellular map.

The simulated data sets generated to test the unsupervised factor analysis algorithm model generation approach to CLSA were used again to test the supervised approach which uses component spectra of polystyrene, RNA and lipid as the model spectra to generate scores for each spectrum in the data set. In the simulated data shown in Figure 5, it is observed that it is possible to identify a trend similar to that seen in Figure 2 for the unsupervised CLSA. For RNA and lipid, the trend matches well the predicted response for concentrations as low as 1:0.1, whereupon it deviates from linearity, falling to zero at a ratio of~1:0.03. However, for polystyrene, although the trends are similar, the results are deviate from the predicted response much earlier than the unsupervised CLSA. This indicates that the identification of the components using a supervised CLSA approach may not be as accurate as the model generation approach shown in Figure 3. Thus, to test this prediction and for comparison, supervised CLSA was carried out on the same cellular data set using polystyrene, RNA and lipid spectra as the cellular components used to generate the scores for CLSA.

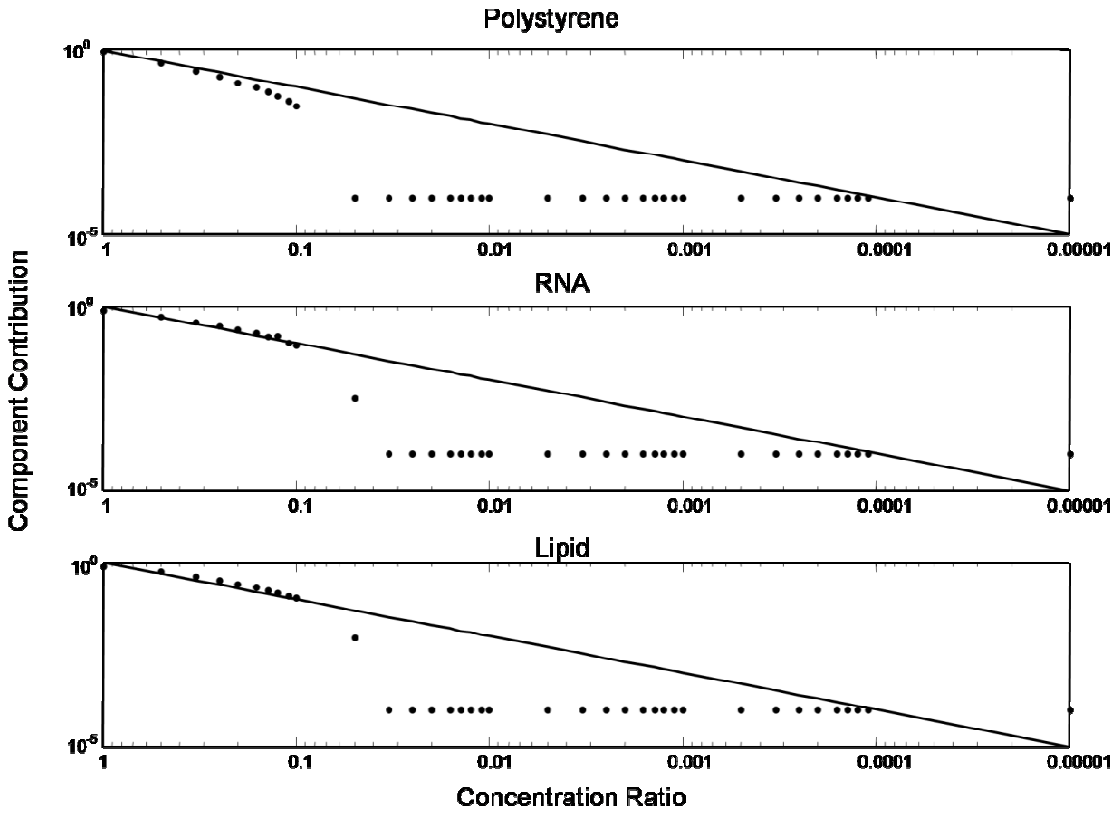


Figure 5. Supervised CLSA of simulated spectral data sets of nano-polystyrene, RNA and lipid. In each graph, either the pure spectrum of polystyrene, RNA or lipid was used to calculate the CLSA score. This score was then plotted against the concentration ratio of pure component spectrum: cellular spectrum used to generate the simulated data set.

Single Cell Data - Supervised CLSA

In order to compare the different CLSA approaches, the cellular Raman data set was screened using three pure component spectra individually, nano-polystyrene, RNA and lipid. The aim was to use these spectra to generate the CLSA scores and thus identify regions of the cell which

correspond to each spectrum, identifying different regions of the cell based on their biochemical composition and also where the nanoparticles were situated.

The spectra and corresponding score maps are shown in Figure 6 A – C. Figure 6A shows a spectrum of polystyrene which was used to screen the map and corresponding visual image of the distribution of nano-polystyrene in the cell. In the image, it is observed that the polystyrene is present in every spectrum in the cell, albeit in differing amounts based on the pixel intensity at each point. This is not consistent with the model generated CLSA above or with previously published data which show the polystyrene to be localised in clusters surrounding the nucleus[39]. However, the regions of high intensity most likely correspond to the areas which contain the nanoparticles.

Similarly this method for assessing the distribution of RNA and lipids in the cell does not quite reproduce the results observed above for CLSA using the unsupervised factor analysis algorithm. Again, it is observed that the distribution of lipid and RNA is throughout the Raman map of the cell, which, while more plausible for lipids, does not make biological sense for the RNA. Therefore, again it must be concluded that the supervised CLSA approach is prone to error, although it is still possible to compare regions of high intensity to the output of the unsupervised CLSA images above. An arbitrary threshold can be applied to the dataset, as is shown for the three component spectra in the right hand panels of Figure 6A-C. Using this method, the spatial distributions of the components matches well that of the unsupervised CLSA. However this threshold is ambiguous and it is not possible to say from the simulated data at what value an accurate representation of the biochemical distribution in the cell is achieved.

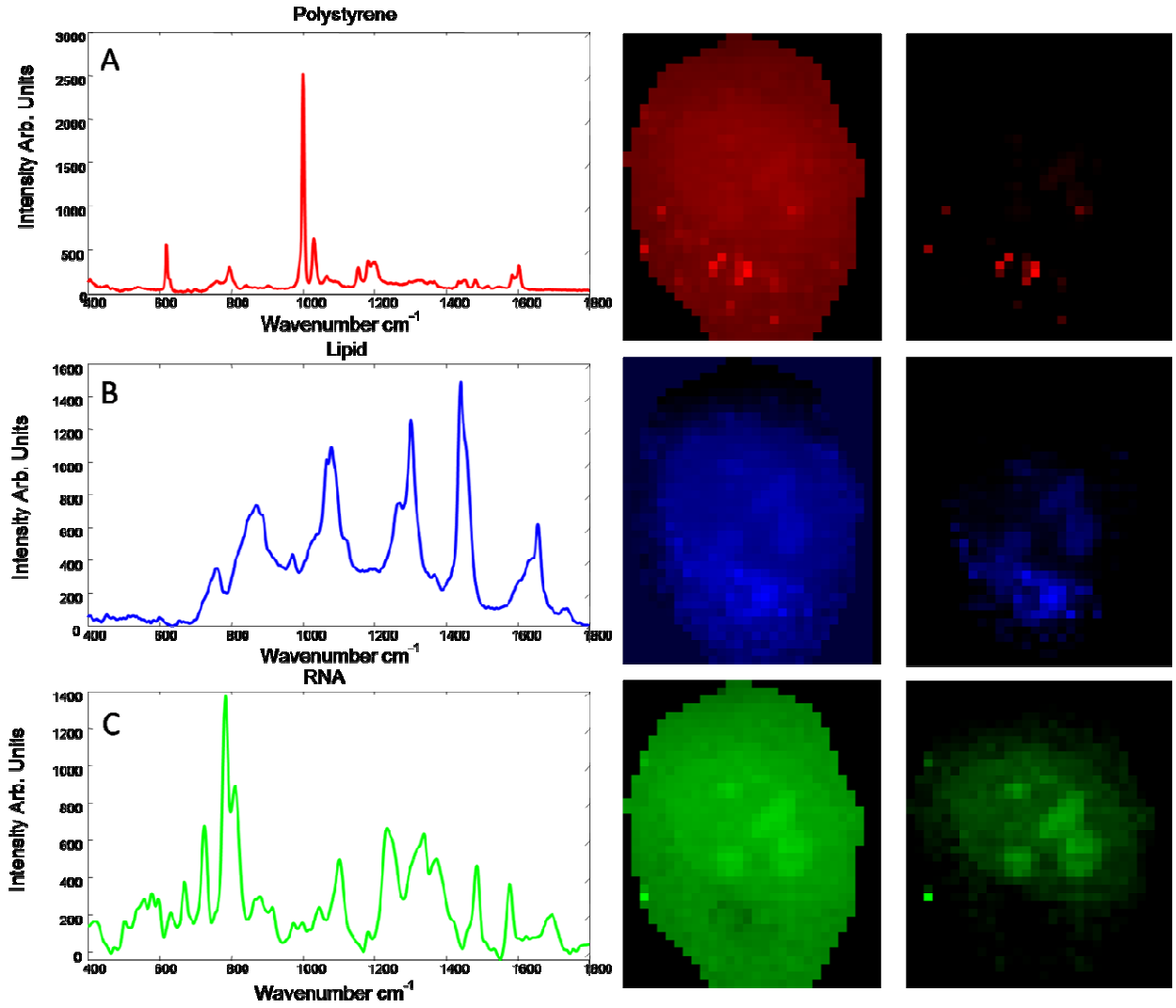


Figure 6: Supervised CLSA using component spectra of polystyrene (A), RNA (B) and (C) 3-sn-phosphatidyl ethanolamine. The spectrum of each pure component is shown on the left of the graph, with the corresponding to non-thresholded data shown in the middle and arbitrarily thresholded data shown on the right.

Simulated data –Spectral Cross Correlation Analysis

The observations in Figure 6 that supervised CLSA contained a high level of error in the Raman images prompted a search for an alternate supervised approach to screening Raman data sets which could be used to unambiguously identify regions of the cell which correspond to the pure component spectrum of interest chosen, be that polystyrene, RNA, lipid or any other spectral signature which may be of interest. A novel technique was thus investigated for the analysis of Raman maps, which uses cross correlation as a method to investigate the presence or absence of a component in a complex Raman data set in a supervised manner. Thus, SCCA was used to screen the same simulated and real data sets for the presence of polystyrene, RNA and lipid for comparison which both methods of CLSA.

Spectral cross correlation analysis (SCCA) was initially investigated using the same simulated data sets that were used to investigate both CLSA approaches. Similar to the supervised CLSA approach, pure component spectra were used to screen each data set for the presence of each in their respective simulated data set. Figure 7 compares the results of the simulated SCCA for each of the different components polystyrene, lipid and RNA. In all cases, a correlation of the SCCA co-efficient and the true concentration ratios is observed, but to varying degrees of accuracy.

For polystyrene, a minimum correlation coefficient value of ~ 0.3 is reached at a concentration ratio of cellular: polystyrene spectrum of ~ 1:0.1. This indicates that at this concentration ratio, the presence of the polystyrene spectral fingerprint cannot be distinguished from the cellular spectrum. Thus, for the practical implications of screening a cell for polystyrene nanoparticles, correlation coefficient values at or below 0.3 represent the cellular peaks which

overlap with characteristic polystyrene peaks and thus values below this are deemed not to be nanoparticles. This hypothesis was tested using a blank Raman map which contained no polystyrene data in (data not shown) and a value of correlation of 0.3125 was determined, which is close to the predicted value in the simulated data sets. This indicates the need to threshold cellular data in order to identify polystyrene nanoparticles in the cell.

A similar performance was observed for both RNA and lipid simulated data sets, where an initial decrease in the correlation coefficient was observed in relation to concentration ratio of pure component: cell spectrum. Again a minimum baseline correlation coefficient was observed for both RNA and lipid simulated SCCA data. Notably, however, this value was different, in both cases higher, than that observed for polystyrene, possibly due to an increased overlap of Raman bands present in the lipid and RNA spectra with cellular Raman bands in comparison to the polystyrene spectrum. In the case of the lipid contribution, the correlation with the predicted response is quantitatively poor even at ratios above 1:0.1. However, this can possibly be explained by lipid contributions already present in the cellular spectrum and/or the relatively broad lipid bands present in the lipid spectrum used.

The next step was to investigate the performance of SCCA in a real Raman data set of the cell. Thus the previous map was screened in a supervised manner to investigate if nano-polystyrene could be identified in the Raman map. Additionally, the lipid spectrum was used to see if the local cell environment could be investigated. Also, as used in the above supervised CLSA, RNA was used to see if a differentiation could be made between the nucleus and nucleolus.

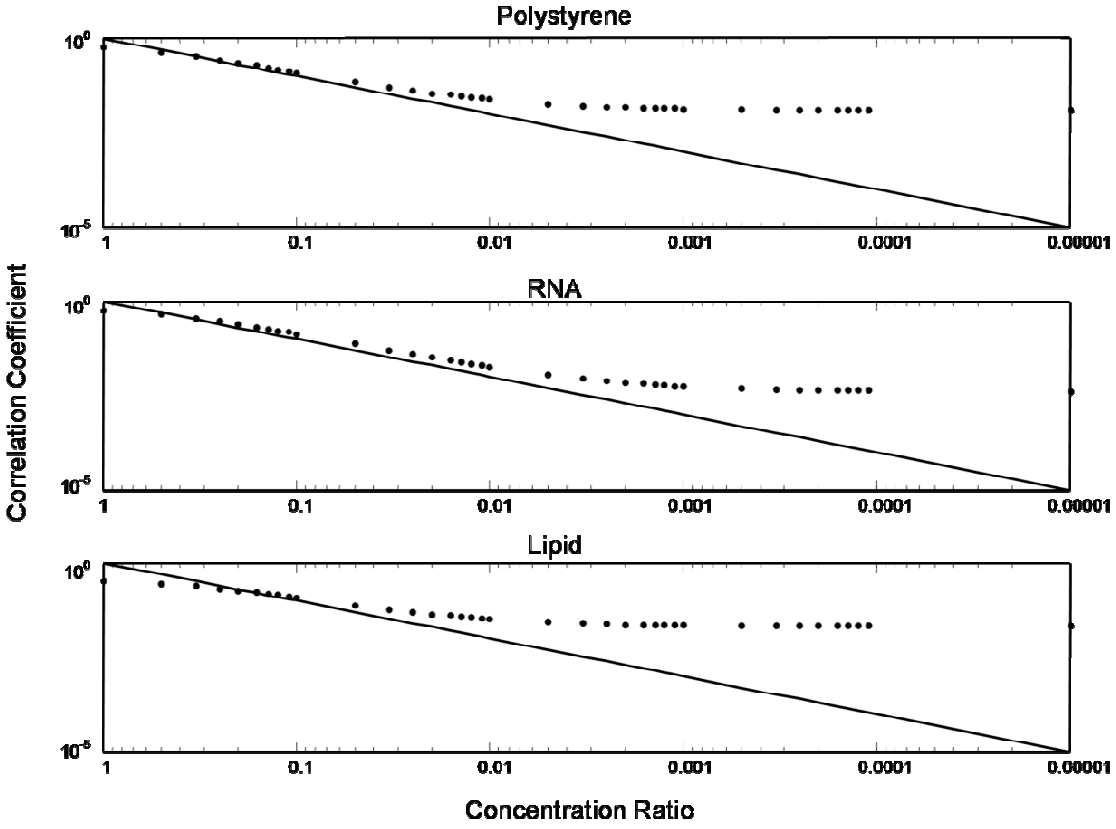


Figure 7. SCCA carried out on simulated data sets containing added polystyrene, RNA and lipid component spectra. In each instance, a pure component spectrum of polystyrene, RNA and lipid was cross correlated against each data set to investigate the performance of the technique. The solid line shows the idealised response.

Single Cell Data –SCCA

SCCA was used to screen the Raman data set for the presence of polystyrene, RNA and lipid distributions. The spectra and correlation maps are shown in Figure 8. In figure 8A, the spectrum of polystyrene is shown in red and the corresponding correlation map is shown adjacent for both thresholded (right) and non-thresholded (left) datasets. This map shows the distribution of polystyrene nanoparticles in the Raman map. Importantly, the threshold which was predicted from the simulated data, or more simply from a cross-correlation of the component spectrum with the raw average cellular spectrum, was applied to the data set and returned a map which

corresponded to the previously observed Raman image from the unsupervised CLSA (Fig 4A). Notably, however, the spectrum is the pure spectrum of polystyrene, rather than a cellular/polystyrene mixture. This result shows the capability for a supervised approach for the unambiguous identification of polystyrene nanoparticles in complex Raman spectroscopic data sets.

Furthermore, to investigate how SCCA can be used to probe the local cellular environment, the lipid spectrum was used to screen the data set (Fig 8B). Again applying a threshold to the data set it is possible to identify regions of the cell which contain a high density of lipids using a supervised approach to Raman analysis. Thus it is possible to investigate the local cell environment to which the nanoparticles are trafficked after 24hrs. This is consistent with the previous K-means cluster analysis [39] which suggests that indeed the nanoparticles are located in a highly lipid rich environment.

As an additional demonstration of the potential of SCCA, a pure RNA spectrum was cross correlated against the data set to see if it was possible to differentiate spectra which corresponded to the nucleolus of the cell and thus differentiate between DNA and RNA rich regions of the cell. Figure 8C shows that it is possible to identify the nucleolus of the cell using cross correlation analysis. It was also observed that a high correlation coefficient was present in regions outside the nucleus. This could possibly correspond to cytoplasmic ribosomal RNA (rRNA) or cytoplasmic messenger RNA (mRNA). Thus a novel approach for extracting complex spectral information from Raman data sets is demonstrated in SCCA.

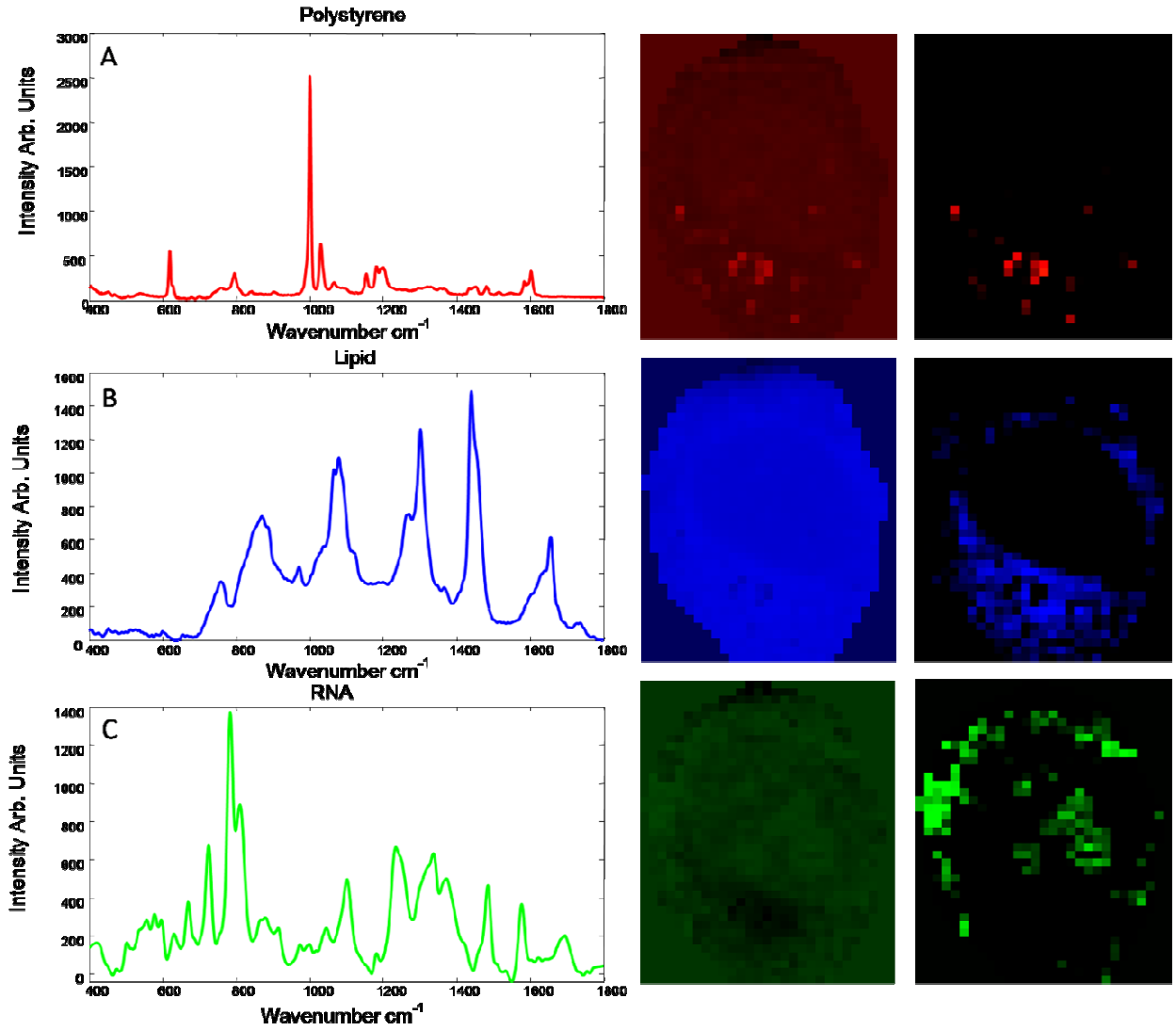


Figure 8: SCCA analysis using component spectra of polystyrene (A), 3-sn-phosphatidyl ethanolamine (B) and RNA (C). The spectrum of each pure component is shown on the left of the figure and the correlation maps for non-thresholded shown in the middle and thresholded on the right.

Discussion

Raman spectroscopy is a powerful tool for the investigation of biological samples. Previous studies have shown the capability of the technique to investigate sub cellular structures and processes which provide Raman images comparable to images observed using wide-field and

confocal fluorescent microscopy[45,46,35,39,47]. Notably, however, Raman spectroscopy is a label free method which provides a visualization of the biochemical make up of a cell without costly and time consuming processing with reagents, and when combined with appropriate analysis methods can provide a wealth of information pertaining to biological processes in the cell. The aim of this paper was thus to investigate two analytical approaches both in an unsupervised and supervised approach and assess their ability to identify polystyrene nanoparticles and biochemical distributions in a single cell Raman map.

Unsupervised CLS analysis is demonstrated to be capable of identifying the presence of nanoparticles in regions of the cell. However, while this method is valuable for identifying distributions in the cell, the model spectra generated in this manner must be further analysed to extract any real biochemical information. Therefore, while the analysis of the simulated dataset in figure 2 indicates that the unsupervised model has a higher accuracy, the model spectra yielded by the unsupervised CLS analysis do not directly compare to the pure component spectra shown in Figure 1 and therefore cannot be used to unambiguously identify the contributing components.

In contrast, employing supervised approaches to the analysis of Raman data sets allows for the spectral array to be screened directly with the nanoparticle or pure biochemical component spectrum of interest. Analysis in this way enables a direct screening of the cellular distribution of a particular component while simultaneously probing the chemical or biochemical environment of the particular location in the cell. CLSA and SCCA are both used in a supervised approach for analysing Raman cellular data sets (Figure 6 and Figure 8). However, unthresholded, both show a degree of error for all three components tested (nano-polystyrene, RNA and Lipids). To correct for this, a threshold can be applied to both CLSA and SCCA.

Importantly, this threshold should not be applied in an arbitrary manner, as this facilitates a loss of information from the dataset. While thresholding for supervised CLSA is arbitrary and subjective, the simulated datasets generated for SCCA provided a good estimation of where this thresholding should take place and in combination with cellular data containing no nanoparticles it was possible to accurately reveal where the nanoparticles were located in the cell. It should be noted that the thresholding level appears to be dependent on the spectral profile of the individual component, as it is dependent on the degree of similarity of the spectrum of the target component with that of the environment. Incorrect correction of spectral background may also add to the threshold. On the other hand the simulated data for supervised CLSA did not provide a threshold value to apply to the dataset and thus was arbitrarily thresholded, which is far from ideal to gain any reliable information about the dataset. Therefore, SCCA provides a more reliable supervised approach for identification of nanoparticles and other biological components when used in combination with a threshold generated by simulated datasets. In addition, quantitative information can be extracted from the simulated data sets, with each of the three approaches showing some level of quantification based on how well the matched the predicted response, with SCCA showing the highest level of sensitivity of the three techniques. SCCA is specifically a supervised approach, as it is necessary to provide the pure component spectrum. However, it is conceivable the technique could be extended to a library of reference spectra which could in turn be screened against the data set in an unsupervised manner.

Conclusions

CLSA and SCCA are shown to be two methods capable of identifying intracellular polystyrene nanoparticles and also to probe the local biochemical environment the nanoparticles are

trafficked to within the cell. CLSA is a relatively straight forward method for analysing spectroscopy data sets. However, SCCA is demonstrated in the simulated data sets to be a more sensitive approach for nanoparticle identification. It is envisaged that both these and other supervised methods will provide analytical approaches which can be used not only as identification methods for other nanoparticles inside cells and detection of resultant biochemical changes, but also to provide alternate analytical approaches to the study of other processes such as chemotherapeutic response of cells to drugs. Additionally the full quantitative nature of these analytical approaches will need to be explored if Raman spectroscopy is to become a routine application in the study of nano-bio interactions and beyond.

Acknowledgements: This research was supported by the Integrated NanoScience Platform, Ireland (INSPIRE), and the National Biophotonics and Imaging Platform (NBIP) Ireland, both funded under the Higher Education Authority PRTLI (Programme for Research in Third Level Institutions) Cycles 4 and 5, co-funded by the Irish Government and the European Union Structural fund.

References

1. S. Dhar and N. Kolishetti, *Proceedings of the National Acadamy of Science*, 2011, **108**, 1850–1855.

2. Y. Wang, Y. Wang, J. Xiang, and K. Yao, *Biomacromolecules*, 2010, **11**, 3531–8.

3. T. Lammers, F. Kiessling, W. E. Hennink, and G. Storm, *Journal of controlled release: official journal of the Controlled Release Society*, 2012, **161**, 175–87.

4. A. Kumar, P. K. Vemula, P. M. Ajayan, and G. John, *Nature materials*, 2008, **7**, 236–41.

5. I. Perelshtein, G. Applerot, N. Perkas, J. Grinblat, and A. Gedanken, *Chemistry (Weinheim an der Bergstrasse, Germany)*, 2012, **18**, 4575–82.

6. H. Yan, H. S. Choe, S. Nam, Y. Hu, S. Das, J. F. Klemic, J. C. Ellenbogen, and C. M. Lieber, *Nature*, 2011, **470**, 240–4.

7. S. S. Kelkar and T. M. Reineke, *Bioconjugate Chemistry*, 2011, **22**, 1879–903.

8. S. Wang, G. Kim, Y.-E. K. Lee, H. J. Hah, M. Ethirajan, R. K. Pandey, and R. Kopelman, *ACS Nano*, 2012.

9. J.-P. Colinge, C.-W. Lee, A. Afzalian, N. D. Akhavan, R. Yan, I. Ferain, P. Razavi, B. O'Neill, A. Blake, M. White, A.-M. Kelleher, B. McCarthy, and R. Murphy, *Nature Nanotechnology*, 2010, **5**, 225–9.

10. T. R. Downs, M. E. Crosby, T. Hu, S. Kumar, A. Sullivan, K. Sarlo, B. Reeder, M. Lynch, M. Wagner, T. Mills, and S. Pfuhler, *Mutation Research*, 2012, **745**, 38–50.

11. M.-F. Song, Y.-S. Li, H. Kasai, and K. Kawai, *Journal of Clinical Biochemistry and Nutrition*, 2012, **50**, 211–6.

12. B. Ziemba, A. Janaszewska, K. Ciepluch, M. Krotewicz, W. a Fogel, D. Appelhans, B. Voit, M. Bryszewska, and B. Klajnert, *Journal of Biomedical Materials Research. Part A*, 2011, **99**, 261–8.

13. R. P. Singh and P. Ramarao, *Toxicology Letters*, 2012, 1–11.

14. X. Zheng, J. Tian, L. Weng, L. Wu, Q. Jin, J. Zhao, and L. Wang, *Nanotechnology*, 2012, **23**, 055102.

15. A. Jaeger, D. G. Weiss, L. Jonas, and R. Kriehuber, *Toxicology*, 2012, **296**, 27–36.

16. G. Hunt and M. Riediker, *Nanotechnology Perceptions*, 2011, **7**, 82–98.

17. H. J. Byrne, I. Lynch, W. H. D. Jong, W. G. Kreyling, S. Loft, M. V. D. Z. Park, M. Riediker, and D. Warheit, 2008, 1–30.

18. P. Sandin, L. W. Fitzpatrick, J. C. Simpson, and K. a Dawson, *ACS Nano*, 2012, **6**, 1513–21.

19. F. Fazlollahi, S. Angelow, N. R. Yacobi, R. Marchelletta, A. S. L. Yu, S. F. Hamm-Alvarez, Z. Borok, K.-J. Kim, and E. D. Crandall, *Nanomedicine: Nanotechnology, Biology, and Medicine*, 2011, **7**, 588–94.

20. E. Jan, S. J. Byrne, M. Cuddihy, A. M. Davies, Y. Volkov, Y. K. Gun'ko, and N. a Kotov, *ACS Nano*, 2008, **2**, 928–38.

21. T. Y. Ohulchanskyy, I. Roy, K.-T. Yong, H. E. Pudavar, and P. N. Prasad, *Wiley Interdisciplinary Reviews. Nanomedicine and Nanobiotechnology*, 2010, **2**, 162–75.

22. G. D. Byrne, M. C. Pitter, J. Zhang, F. H. Falcone, S. Stolnik, and M. G. Somekh, *Journal of Microscopy*, 2008, **231**, 168–79.

23. J. Contreras, J. Xie, Y. J. Chen, H. Pei, G. Zhang, C. L. Fraser, and S. F. Hamm-Alvarez, *ACS Nano*, 2010, **4**, 2735–2747.

24. T. Tenuta, M. P. Monopoli, J. Kim, A. Salvati, K. a Dawson, P. Sandin, and I. Lynch, *PloS one*, 2011, **6**, e25556.

25. G. Wang, A. S. Stender, W. Sun, and N. Fang, *Analyst*, 2010, **135**, 215–21.

26. J. Zhou, C. Leuschner, C. Kumar, J. Hormes, and W. O. Soboyejo, *Materials Science and Engineering: C*, 2006, **26**, 1451–1455.

27. A. M. Schrand, J. J. Schlager, L. Dai, and S. M. Hussain, *Nature Protocols*, 2010, **5**, 744–57.

28. K. Shapero, F. Fenaroli, I. Lynch, D. C. Cottell, A. Salvati, and K. a Dawson, *Molecular BioSystems*, 2011, **7**, 371–8.

29. M. Davoren, E. Herzog, A. Casey, B. Cottineau, G. Chambers, H. J. Byrne, and F. M. Lyng, *Toxicology In vitro: an International Journal Published in Association with BIBRA*, 2007, **21**, 438–48.

30. F. M. Lyng, E. O. Faoláin, J. Conroy, a D. Meade, P. Knief, B. Duffy, M. B. Hunter, J. M. Byrne, P. Kelehan, and H. J. Byrne, *Experimental and Molecular Pathology*, 2007, **82**, 121–9.

31. K. M. Ostrowska, A. Malkin, A. Meade, J. O'Leary, C. Martin, C. Spillane, H. J. Byrne, and F. M. Lyng, *Analyst*, 2010, **135**, 3087–93.

32. H. Nawaz, F. Bonnier, P. Knief, O. Howe, F. M. Lyng, A. D. Meade, and H. J. Byrne, *Analyst*, 2010, **135**, 3070–6.

33. H. Nawaz, F. Bonnier, A. D. Meade, F. M. Lyng, and H. J. Byrne, *Analyst*, 2011, **136**, 2450–63.

34. F. Bonnier, P. Knief, B. Lim, a D. Meade, J. Dorney, K. Bhattacharya, F. M. Lyng, and H. J. Byrne, *Analyst*, 2010, **135**, 3169–77.

35. F. Bonnier, a D. Meade, S. Merzha, P. Knief, K. Bhattacharya, F. M. Lyng, and H. J. Byrne, *Analyst*, 2010, **135**, 1697–703.

36. P. Knief, C. Clarke, E. Herzog, M. Davoren, F. M. Lyng, A. D. Meade, and H. J. Byrne, *Analyst*, 2009, **134**, 1182–91.

37. K. A. M. R. A. N. Badizadegan, N. Yoshizawa, C. Boone, and M. I. S. Feld, *Applied Spectroscopy*, 2002, **56**, 150–154.

38. J. Kneipp, H. Kneipp, M. McLaughlin, D. Brown, and K. Kneipp, *Nano Letters*, 2006, **6**, 2225–31.

39. J. Dorney, F. Bonnier, A. Garcia, A. Casey, G. Chambers, and H. J. Byrne, *Analyst*, 2012, **137**, 1111–9.

40. M. Hedegaard, C. Matthäus, S. Hassing, C. Krafft, M. Diem, and J. Popp, *Theoretical Chemistry Accounts*, 2011, **130**, 1249–1260.

41. M. MiljkoviÄ‡, T. Chernenko, and M. Romeo, *Analyst*, 2010, 2002–2013.

42. M. Miljković, T. Chernenko, M. J. Romeo, B. Bird, C. Matthäus, and M. Diem, *Analyst*, 2010, **135**, 2002–13.

43. F. Bonnier and H. J. Byrne, *Analyst*, 2012, **137**, 322–32.

44. C. L. Zavaleta, B. R. Smith, I. Walton, W. Doering, G. Davis, B. Shojaei, M. J. Natan, and S. S. Gambhir, *Proceedings of the National Academy of Sciences of the United States of America*, 2009, **106**, 13511–6.

45. C. Matthäus, T. Chernenko, J. a Newmark, C. M. Warner, and M. Diem, *Biophysical Journal*, 2007, **93**, 668–73.

46. M. Miljković, T. Chernenko, M. J. Romeo, B. Bird, C. Matthäus, and M. Diem, *Analyst*, 2010, **135**, 2002–13.

47. K. Klein, A. M. Gigler, T. Aschenbrenner, R. Monetti, W. Bunk, F. Jamitzky, G. Morfill, R. W. Stark, and J. Schlegel, *Biophysical Journal*, 2012, **102**, 360–8.